



icbi

THE INNOVATION CENTER FOR BIOMEDICAL INFORMATICS

# Data analysis in G-DOC *Plus* using Variant Search tool

*Innovation Center for Biomedical Informatics (ICBI)*

*Georgetown University*

*Feb 2015*

# Introduction

- Sequence variations within genome lead to phenotypic differences, including predisposition to genetic diseases and response to environmental factors. Data generated from whole genome sequencing (WGS) of normal or diseased tissues from individuals/cell lines are used to identify germline or somatic variants respectively. Analysis of such vast amount of genomic data can be made tractable and meaningful through variant search, annotation and analyses tools.
- The G- DOC *Plus* “Variant Search” tool enables several functionalities including searching for specific variants, eg., functional variants, their stratification and downstream analysis of selected variant genes for pathway profiling or Cancer Gene Index based network analysis.
- Once variant search is complete, results can be saved in G-DOC *Plus* for further downstream analysis that includes profiling of genes in the saved list based on functional pathways, as well as to obtain an idea of networked relationship between individual genes derived from Variant search analysis.

# Introduction - 2

- Users of this tool are expected to have knowledge of clinical and biological analysis of variant data. The tool can be used for exploratory purposes for generating hypothesis or to test an existing hypothesis.
- In this tutorial, we will go through 3 different examples, starting from a simple Variant search, defining criteria for search, and finally use results from search for downstream analysis. The Variant Search tool will be used to analyze a public data set on breast cancer cell lines obtained from Complete Genomics BRC\_CG\_XXXX\_01.

*Note: please ensure that your computer passes systems requirement check (<https://gdoc.georgetown.edu/gdoc/home/requirementCheck>). Failure to do so may prevent successful use of tool.*

Log into G-DOC *Plus*  
<https://gdoc.georgetown.edu>



Home

Studies

Lists

Analyses

Groups

Notifications

Study Options ▾

Help



kb472 ▾

# G-DOC Plus Launch Pad!

Welcome back, your last login was Mon Feb 9, 2015. You can check if you have been granted access to new lists or analyses since your last login



Welcome! The G-DOC Plus Launch Pad is your one-stop resource for learning more about G-DOC and getting started on the platform.



Studies



Lists



Groups

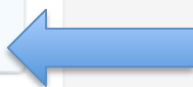


Notifications 0

## It All Starts Here!

G-DOC has over seventy studies, We know this can be overwhelming! Let us guide you to choose the study that is relevant for your research.

[Let's Go!](#)




# Selections


- "What's your area of interest" : Precision medicine - it should go to the workflow page.
- Select "data collection" : BREAST CANCER
- Select a study : BRC\_CG\_XXXX\_01 (Note: If you click on "More", you will be able to see the complete description of the study)
- Select the tool: Variant Search

## What's your area of interest?


G-DOC Plus has three overlapping entry points for the user based on their interests. Choose your area of interest to launch the workflow.



**Precision Medicine**  
Patients' molecular diagnostics and clinical data.



**Translational Research**  
Analytic tools and workflows to enable discovery.



**Population Genetics**  
Race-based, genomic reporting and comparison.

## Precision Medicine

What type of data collection do you want to analyze?



### BREAST CANCER

2 studies   93 samples   4 biospecimen

## Precision Medicine

Study Selected! BREAST CANCER → BRC\_CG\_XXXX\_01



Based upon the study you picked, here is a list of tools you can use:

Search

- Variant Search

## Precision Medicine

Great! What breast cancer study would you like



### BRC\_CG\_XXXX\_01

**Title:** Complete Genomics Breast Cancer dataset

**Data Type Details:** CLINIC,WGS

**Abstract:** Sample dataset from Complete Genomics Cancer Sequencing Service. It contains matched tumor and normal cell line sequence data for two patients with breast cancer. Collections were drawn from ATCC. Samples were sequenced in an average

4 samples   4 biospecimen

[More>>](#)

### BRC\_IMAGING\_TCIA\_01

**Title:** Breast-Diagnosis imaging collection from TCIA

**Data Type Details:** IMAGING,CLINIC

**Abstract:** The Breast-Diagnosis collection from TCIA contains cases that are high-risk normals, DCIS, florid and lobular carcinomas. Each case has 3 or more distinct MR pulse sequences from a Philips 1.5 T (breast) sequences are labeled T2, STIR and

89 samples   0 biospecimen

[More>>](#)

**Tip:** Click "More" to see complete description of study. Click "Select study"

# Example 1: Finding all variants in BRCA1

## Search Sequence Variations

Current Study: BRC\_CG\_XXXX\_01 [change study?](#)

### Genes

BRCA1 x

### Chromosome

Select a chromosome ▾

### Functional Location

Functional Location

### Exonic Function

☒ Non-Synonymous

### Variant Inclusions

☐ Known

☐ Novel

### Sample IDs (max 2)

Enter Sample ID

### Variants

Chromosome	Start	Reference	Alleles	XRefs	Gene	Functional Location	Exonic Function	REF_FREQ	ALT_FREQ	MISSING
chr17	41194885	T	C	rs7223952	BRCA1	Upstream of gene		0.5	0.5	
chr17	41195025	T	A	rs11659028	BRCA1	Upstream of gene		0.5	0.5	
chr17	41195711	G	C	rs8176323	BRCA1	Upstream of gene		0.5	0.5	
chr17	41195773	A	G	rs8176322	BRCA1	Upstream of gene		0.5	0.5	
chr17	41196408	G	A	rs12516	BRCA1	3' UTR		0.5	0.5	
chr17	41196821	CTT	C	rs139756895	BRCA1	3' UTR			0.125	0.875
chr17	41197274	C	A	rs8176318	BRCA1	3' UTR		0.5	0.5	
chr17	41198621	A	G	rs8176314	BRCA1	Intronic		0.5	0.5	
chr17	41198774	C	CA	.	BRCA1	Intronic			0.125	0.875
chr17	41199913	T	C	rs8176310	BRCA1	Intronic		0.5	0.5	
chr17	41200109	T	C	rs4793190	BRCA1	Intronic		0.5	0.5	
chr17	41200537	T	C	rs4792972	BRCA1	Intronic		0.375	0.375	0.25
chr17	41200704	GT	G	rs145809091	BRCA1	Intronic		0.5	0.375	0.125
chr17	41201702	C	T	rs3092988	BRCA1	Intronic		0.5	0.5	
chr17	41202632	AAAG	A, AAAGA	.	BRCA1	Intronic		0.375	0.25, 0.125	0.25
chr17	41202688	G	A	rs8070179	BRCA1	Intronic		0.5	0.5	
chr17	41202822	G	A	.	BRCA1	Intronic		0.625	0.375	

# Column Definitions

- **Chromosome**- on which chromosome the variant gene is present,
- **start position** of the reference allele,
- **Reference** and **Alternate** alleles,
- **XRef**- provides rsIDs from dbSNP where available,
- **Gene** - HUGO gene symbol,
- **Functional location** as to whether the variation occurs on exons, UTR, etc in the gene,
- **Exonic function**- type of the variant as to a deletion, frameshift, stop codon gain or loss etc,
- **Ref freq, Alt freq and Missing freq**: frequencies of reference and alternate alleles, and instances of missing frequencies in case of deletions/insertions, respectively.
- Above results can be saved in GDOC by clicking on **Save gene list** or exported to your drive by **Export results**.
- *Notes:*
  - You can add more genes by typing a gene symbol with cursor placed after the **X** sign. To remove any of the selected genes, click on the **X** sign following the name of the gene.
  - Multiple selections in the **Functional location** option is an “OR” function. Eg: if you select Coding genes and 3’ UTR, it will show all variants present in either coding genes or 3’UTR prime regions

# Example 2

- **Example 2.** Let us find all functionally impacting variants in all genes located in Chromosome 4, that will give rise to truncated proteins.
- Make the following selections:
  - **Chromosome:** *chromosome 4*,
  - **Functional Location:** *Coding sequence*,
  - **Exonic Function:** *Non Synonymous -> Premature stop*.

**Search Sequence Variations**

Current Study: BRC\_CG\_XXXX\_01 [change study?](#)

**Genes**

Enter Gene Symbol

**Chromosome**

4

**Functional Location**

Coding sequence X

**Exonic Function**

☒ Non-Synonymous

- ☐ Loss of stop
- ☒ Premature stop
- ☐ Insertion
- ☐ Deletion
- ☐ Frameshift
- ☐ Substitution
- ☐ Possible splice variant
- ☐ Possible 5' splice variant
- ☐ Other

**Variant Inclusions**

- ☐ Known
- ☐ Novel

**Sample IDs (max 2)**

Enter Sample ID

**Variants**

Chromosome	Start	Reference	Alleles	XRefs	Gene	Functional Location	Exonic Function	REF_FREQ	ALT_FREQ	MISSING
chr4	1087487	G	A	rs60035268	RNF212	Coding sequence	Premature stop	0.375	0.625	
chr4	9784217	G	A	.	DRD5	Intronic	Substitution	0.125	0.125	0.75
chr4	26744202	C	T	cosmic:32224 rs141345574	TBC1D19	Coding sequence	Premature stop	0.75	0.25	
chr4	76507104	G	C	cosmic:596	CDKL2	Coding sequence	Premature stop	0.875	0.125	
chr4	100064326	A	T	rs3919370	ADH4	Coding sequence	Premature stop	0.75	0.25	
chr4	100203447	A	C	rs2276332	LOC1005	Intronic	Substitution	0.875	0.125	
chr4	109541499	C	T	rs7677415	ADH1A	Coding sequence	Premature stop	0.125	0.875	
chr4					LOC1005	Intronic	Substitution			
chr4					LOC2854	Coding sequence	Premature stop			
chr4					Non-coding L	Substitution				

Annotated list of variants

Save gene list Export results Page 1 of 1 50 View 1 - 7 of 7



# Example 3

- **Example 3.** Chromosome 8 abnormalities are often reported in breast cancer.
  - What are the genes that might be affected by novel deletions in chromosome 8 with potential impact on protein function, and
  - what major pathways might these genes be involved?
  - Is there a networked relationship between one or more of these impacted genes?
  - How to view this gene list ?
  - How to export gene list

- Make the following selections:
  - **Chromosome:** *chromosome 8*,
  - **Functional Location:** *Coding sequence*,
  - **Exonic Function:** *Non Synonymous -> Deletion*
  - **Variant inclusion :** *Novel*
- **Save gene list** , give a name BRC-8-CDS-Del\_Novel.

## Search Sequence Variations

Current Study: BRC\_CG\_XXXX\_01 [change study?](#)

**Genes**

Enter Gene Symbol

**Chromosome**

8

**Functional Location**

Coding sequence X

**Exonic Function**

☒ Non-Synonymous

- ☐ Loss of stop
- ☐ Premature stop
- ☐ Insertion
- ☒ Deletion
- ☐ Frameshift
- ☐ Substitution
- ☐ Possible splice variant
- ☐ Possible 5' splice variant
- ☐ Other

**Variant Inclusions**

☐ Known

☒ Novel

**Sample IDs (max 2)**

Enter Sample ID

Chromosome	Start	Referenc	Alleles	XRefs	Gene	Functional Lo	Exonic Function	REF_FREQ	ALT_FREQ	MISSING_
chr8	10480173	GCGGC	G	.	RP1L1	Coding seque	Deletion Frameshift	0.625	0.125	0.25
chr8	23186052	ACCG	A	.	LOXL2	Coding seque	Deletion	0.875	0.125	
chr8	77765302	CCTC	C	.	ZFH4	Coding seque	Deletion	0.625	0.25	0.125
chr8	103573010	CTGCAAC	C	.	ODF1	Coding seque	Deletion	0.125	0.125	0.75
chr8	145113143	TCCC	T	.	OPLAH	Coding seque	Deletion Possible 5' splice	0.875	0.125	
chr8	145625537	CC	G	.	CPSF1 MIR1234	Coding seque Non-coding l Intronic	Deletion Frameshift	0.25	0.125	0.625
chr8	145756159	GC	G	.	ARHGAP5	Coding seque	Deletion Frameshift	0.875	0.125	

Save gene list Export results Page 1 of 1 50 View 1 - 7 of 7

- Once the gene list is saved, click on **Lists** on the Top panel. You will see the gene list grouped under the study name (BRC\_CG\_XXXX\_01)
- On the right side there will be several icons which indicate further downstream analysis. Click on the green icon which indicates **Enrich Gene List** analysis

## Lists

Below are all G-DOC Plus saved Lists. You can view, modify, upload, export and share Lists with other groups.

You can use the filter below to search for a specific list. You can also use the tool Panel on the right to manipulate data in the saved lists.

Filter

☐ Check all for deletion Delete List(s) Upload

Tags	Study	Lists
	AD_BLALOCK_2011_01	
	ALL_NORDLUND_2013_01	
	BRC_BUFFA_2011_01	
	BRC_CG_XXXX_01	
gene	BRC_CG_XXXX_01	<div> <input type="checkbox"/> BRC-8-CDS-Del_N (8 items)           <span>3:49 2/10/2015</span> </div> <div>             Studies: BRC_CG_XXXX_01              Tags: gene           </div>

- Results of **pathway enrichment** (using Reactome Pathways) is shown below

## Pathway Enrichment Results

Pathway	p-value	Overlapping Genes
Inactivation of Cdc42 and Rac	$5.483 \times 10^{-3}$	ARHGAP39
Glutathione synthesis and recycling	$6.699 \times 10^{-3}$	OPLAH
Processing of Intronless Pre-mRNAs	$8.521 \times 10^{-3}$	CPSF1
Post-Elongation Processing of Intronless pre-mRNA	$1.397 \times 10^{-2}$	CPSF1
Processing of Capped Intronless Pre-mRNA	$1.397 \times 10^{-2}$	CPSF1
Glutathione conjugation	$1.518 \times 10^{-2}$	OPLAH
Signaling by Robo receptor	$1.941 \times 10^{-2}$	ARHGAP39
Post-Elongation Processing of Intron-Containing pre-mRNA	$2.061 \times 10^{-2}$	CPSF1
mRNA 3'-end processing	$2.061 \times 10^{-2}$	CPSF1
Transport of Mature mRNA Derived from an Intronless Transcript	$2.121 \times 10^{-2}$	CPSF1
Transport of Mature mRNAs Derived from Intronless Transcripts	$2.181 \times 10^{-2}$	CPSF1
Cleavage of Growing Transcript in the Termination Region	$2.602 \times 10^{-2}$	CPSF1
Post-Elongation Processing of the Transcript	$2.602 \times 10^{-2}$	CPSF1
RNA Polymerase II Transcription Termination	$2.602 \times 10^{-2}$	CPSF1
Rho GTPase cycle	$2.662 \times 10^{-2}$	ARHGAP39
Signaling by Rho GTPases	$2.662 \times 10^{-2}$	ARHGAP39
Gene Expression	$3.140 \times 10^{-2}$	CPSF1

Export results

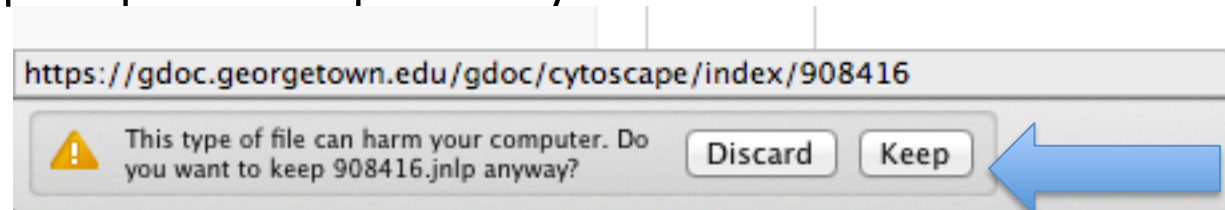
Page 1 of 1 50

View 1 - 27 of 27

- To **check if there is a networked relationship between the genes** in saved gene list, once again go to **Lists** on the top panel.
- Select the red/blue icon on the right side which indicates **View Cancer-Gene Index network**

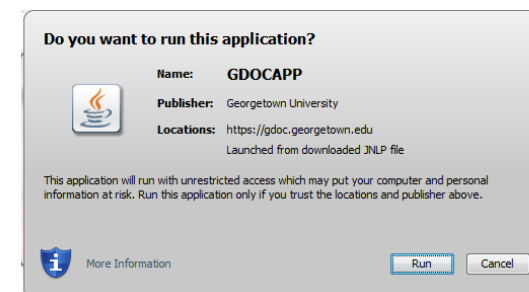


- The above action will prompt a save option on your hard drive

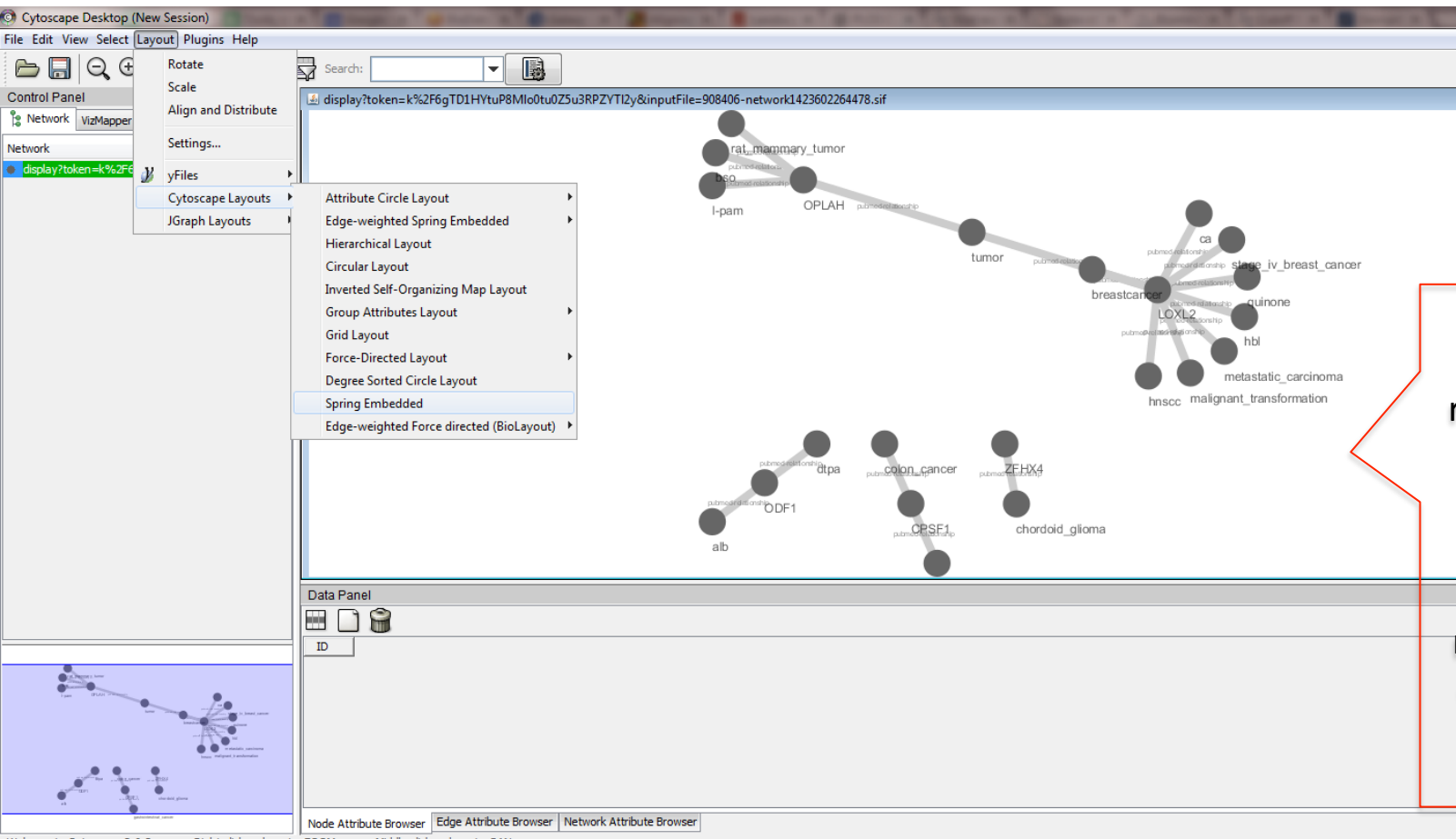


- Save file and open it. As you open it, the application will ask for permission to run Cytoscape (Note: *this may take a few seconds*)
- Click on Run

*Note: If the above does not work for mac books, please check security settings*



- This will open a new page with Cytoscape application showing **networked genes and disease culled from published reports**
- Go to Layout -> Cytoscape Layouts -> Spring Embedded (*if you are familiar with cytoscape, you are free to choose any other layout of your choice*)



Networked  
relation between  
genes on the  
saved gene list  
and annotated  
diseases and  
molecules based  
on Cancer Gene  
Index Network

- To **view the list of genes** in the saved gene list, once again go to **Lists** on the top panel.
- Select the middle icon on the right side which indicates **Export list**. You will now see a text file downloaded to your system



- You can open the exported file to view the list of genes



# Appendix

- **Functional Location:** Physical location of Variants on the chromosome.
  - Functional location filter options: coding sequence, intronic, downstream of gene, upstream of gene, 5' UTR, 3' UTR, non-coding UTR, intergenic
- **Coding gene:** Within a gene which codes for a protein. Within this region are search terms for Intron, downstream of the specified gene, upstream of the specified gene, in the 5'untranslated region, 3'untranslated region.
- **Non-coding region:** Region that does not cover coding genes, it's flanking regions and UTR.
- **Intergenic:** Excludes Coding regions and non-coding UTRs.
- **Non-Synonymous changes**
  - **Exonic Functions:** Includes regions covered only exons of a gene coding for a protein where variation has resulted in non-synonymous changes.
  - Exonic function filter options: Loss of stop (structurally impacted protein), Premature stop (truncated protein), Insertion, Deletion, Frameshift, Substitution, Possible splice variant, Possible 5' splice variant, other



# General tips

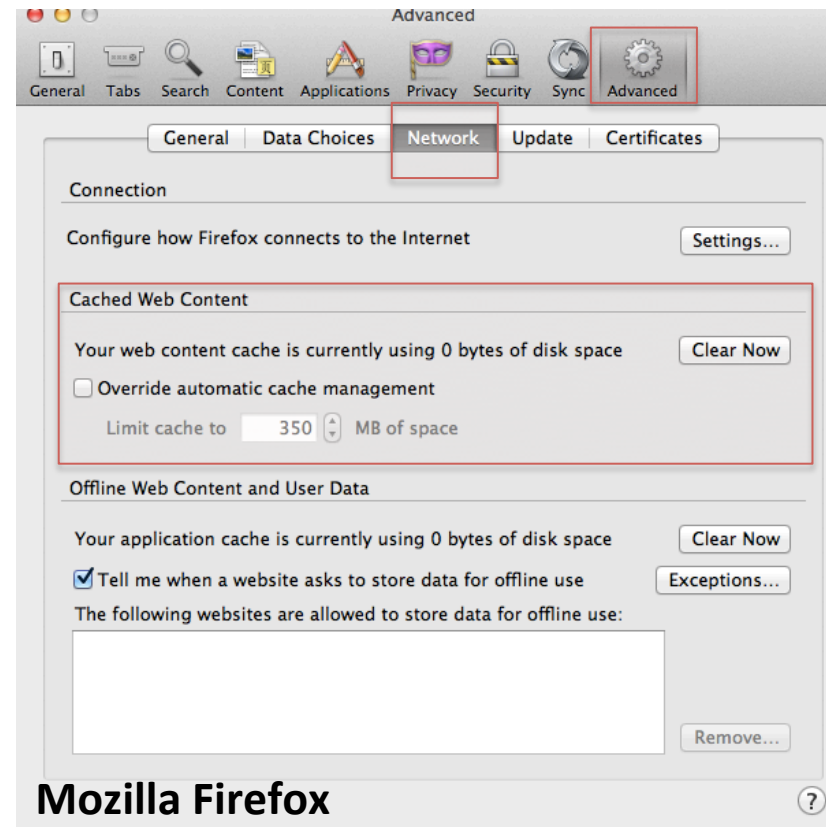
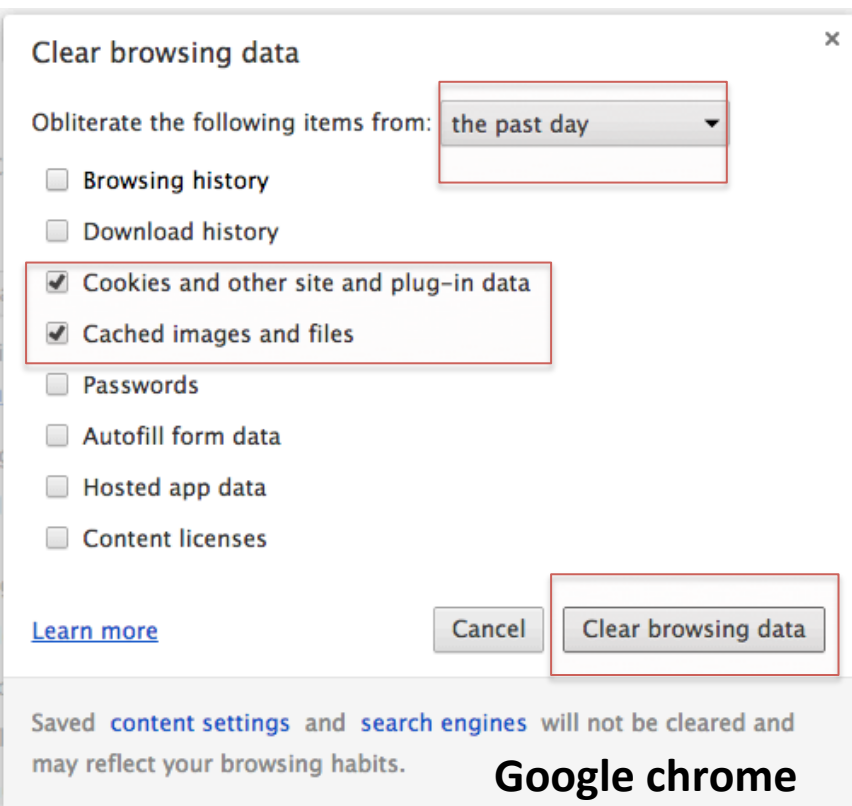
- G-DOC *Plus* works best if you don't use the **back** button in the web browser repeatedly.

Once you select a study, most tools will be easily available from the the top menu bar inside G-DOC *Plus*.

- The Pathway enrichment and the Lists tool may sometimes take a few seconds longer to execute than other tools (since they are directly connecting to the server every time). Your patience is highly appreciated.

# Clearing cache

- If the G-DOC web page does not respond after several seconds, try:
  - refreshing the page.
  - Log out and log back in, and try again
  - If the above two do not work, its possible that your web browser cache may need to be cleared
    - For Google chrome, go to **Settings** -> **Show Advanced Settings** -> Under “Privacy”, select **Clear Browsing data**
    - For Mozilla Firefox, go to **Preferences** -> **Advanced** -> **Network** -> Under “Cached Web Content” -> **Clear now**



- We are working hard to improve G-DOC *Plus*. Please feel free to email your questions and comments (no homework questions please) to us at :[gdoc-help@georgetown.edu](mailto:gdoc-help@georgetown.edu)